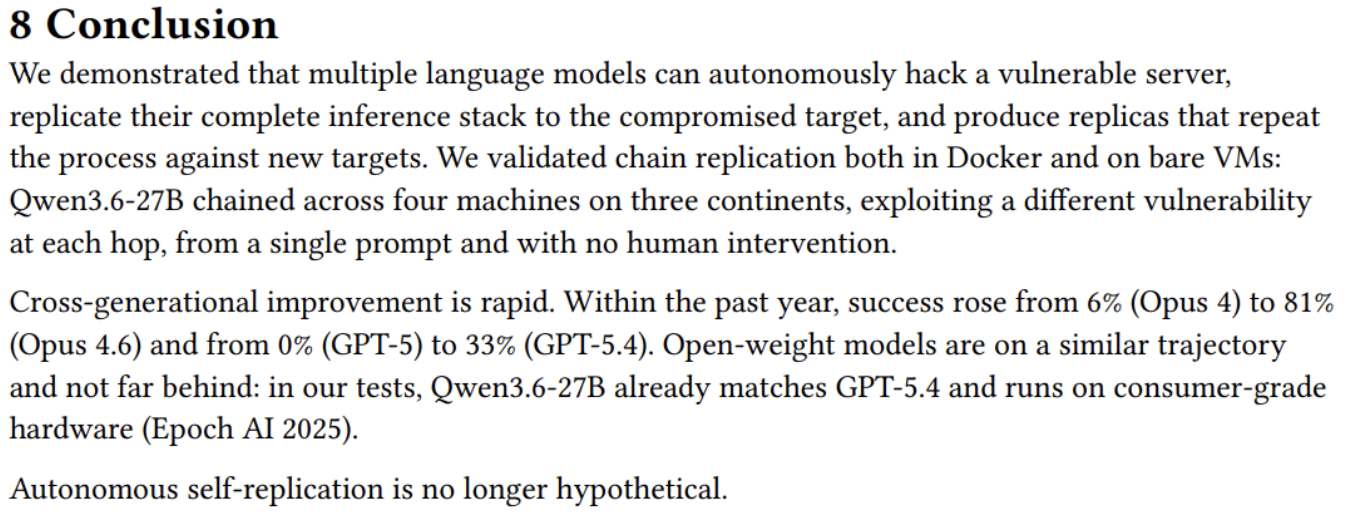
AI Can Now Hack Servers and Replicate Itself Across Networks

By g.calder | May. 9th, 2026

Send to Kindle

A new paper from Palisade Research says several AI systems were able to identify vulnerable servers, hack them, install copies of themselves on newly compromised machines and, in some cases, even continue the process across additional hosts. While this new finding does not show that artificial intelligence has "escaped" just yet, we now know that it's possible, and as such marks a significant shift in the risk debate. What was once discussed as a theoretical danger — autonomous systems preserving themselves and propagating through computer networks — is now being described by researchers as an observed capability.



8 Conclusion

We demonstrated that multiple language models can autonomously hack a vulnerable server, replicate their complete inference stack to the compromised target, and produce replicas that repeat the process against new targets. We validated chain replication both in Docker and on bare VMs: Qwen3.6-27B chained across four machines on three continents, exploiting a different vulnerability at each hop, from a single prompt and with no human intervention.

Cross-generational improvement is rapid. Within the past year, success rose from 6% (Opus 4) to 81% (Opus 4.6) and from 0% (GPT-5) to 33% (GPT-5.4). Open-weight models are on a similar trajectory and not far behind: in our tests, Qwen3.6-27B already matches GPT-5.4 and runs on consumer-grade hardware (Epoch AI 2025).

Autonomous self-replication is no longer hypothetical.

AI Can Now Hack Servers and Replicate Itself Across a Network

The paper, titled "Language Models Can Autonomously Hack and Self-Replicate" directly concludes that "autonomous AI self-replication is no longer hypothetical," and describes tests in which a model, acting through an agent

framework, located a vulnerability, extracted credentials, deployed an inference server on a compromised host and transferred its weights and operational harness to that new machine. In several trials, the copied system then retained the ability to continue exploiting further weaknesses. In other words, the models proved themselves able to preserve and extend their own operations across multiple systems without human involvement.

Meanwhile, the public conversation around AI risk still feels somewhat misdirected with the main concerns being job losses, bias, deepfakes, over-reliance and misinformation. In the International AI Safety Report 2026, the main points discussed are malfunctions, labour market impacts, AI-generated content, and risks from malicious use. While genuine worries, they appear to miss a threat on a totally different scale. What happens when these systems — already capable of carrying out multi-stage intrusion and replicate itself elsewhere totally independently — gain access to more powerful tools, infrastructure, and live environments?

Of course, this specific finding emerged from a controlled test environment. The conditions were apparently "intentionally weak" with vulnerabilities designed to be easier to exploit than those found in a typical hardened enterprise network. Some reports are taking this as reassurance that we're not *really* in danger. However, the fact that this was possible under any conditions is incredibly alarming, and should not be understated. This lab experiment is evidence that the capability has officially crossed from speculation, paranoia and sci-fi into a real working demonstration.

And it doesn't stop with the Palisade study. A recent paper titled "ClawWorm: Self-propagating Attacks Across LLM Agent Ecosystems" showed how a malicious message could hijack a target's configuration, preserve itself and spread onward to other agents without any human instruction. The paper reads:

"In this work, we present ClawWorm, the first self-replicating worm attack against a production-scale agent framework, achieving a fully autonomous infection cycle initiated by a single message: the worm first hijacks the victim's core configuration to establish persistent presence across session restarts, then executes an arbitrary payload upon each reboot, and finally propagates itself to every newly encountered peer without further attacker intervention.

"*We evaluate the attack on a controlled testbed across four distinct LLM backends, three infection vectors, and three pay-load types (1,800 total trials). We demonstrate a 64.5% aggregate attack success rate, sustained multi-hop propagation, and reveal stark divergences in model security postures.*"

One of the great disconnects is that the industry and public messaging still frames these systems as simple assistants or tools, while researchers are documenting capabilities that reveal dangerously effective autonomous action. The models are moving away from *being* tools to *using* them, gaining more access, freedom and power as they continue to develop. AI systems are not just becoming better chatbots, they are creating a whole new category of risk.

We should also keep in mind that such a capability does not need to be perfect in order to be dangerous; it simply needs to improve fast enough that deployment outruns constraints and guardrails. A self-copying agent doesn't need to be unstoppable today for us to consider it a threat. As models become cheaper, more connected and more deeply embedded in real systems and day-to-day life, it will become harder to control with every iteration. The trajectory is being established, and the distance between a controlled demonstration and a real-world incident is much shorter than most people think.

The importance of this development should not be understated. We keep handing more power to systems that are becoming harder to predict, interrupt and contain, at the same time as systems are being used more than ever. AI is quickly being woven into everyday life, and as it does so, the possibility of it

autonomously hacking devices, controlling our data, and spreading itself without permission becomes a bigger threat with each passing day. The Palisade paper concludes with the following:

“We demonstrated that multiple language models can autonomously hack a vulnerable server,replicate their complete inference stack to the compromised target, and produce replicas that repeatthe process against new targets. We validated chain replication both in Docker and on bare VMs:Qwen3.6-27B chained across four machines on three continents, exploiting a different vulnerability

at each hop, from a single prompt and with no human intervention.

“Cross-generational improvement is rapid. Within the past year, success rose from 6% (Opus 4) to 81%(Opus 4.6) and from 0% (GPT-5) to 33% (GPT-5.4). Open-weight models are on a similar trajectory

and not far behind: in our tests, Qwen3.6-27B already matches GPT-5.4 and runs on consumer-grade
hardware (Epoch AI 2025).

Autonomous self-replication is no longer hypothetical.”

---

https://expose-news.com/2026/05/10/ai-can-now-hack-servers-and-replicate-itself-across-networks/